

# You Share, I Share: Network Effects and Economic Incentives in P2P File-Sharing Systems

Mahyar Salek\*

Shahin Shayandeh<sup>†</sup>

David Kempe<sup>‡</sup>

## Abstract

We study the interaction between network effects and external incentives on file sharing behavior in Peer-to-Peer (P2P) networks. Many current or envisioned P2P networks reward individuals for sharing files, via financial incentives or social recognition. Peers weigh this reward against the cost of sharing incurred when others download the shared file. As a result, if other nearby nodes share files as well, the cost to an individual node decreases. Such positive network sharing effects can be expected to increase the rate of peers who share files.

In this paper, we formulate a natural model for the network effects of sharing behavior, which we term the “demand model.” We prove that the model has desirable diminishing returns properties, meaning that the network benefit of increasing payments decreases when the payments are already high. This result holds quite generally, for submodular objective functions on the part of the network operator.

In fact, we show a stronger result: the demand model leads to a “coverage process,” meaning that there is a distribution over graphs such that reachability under this distribution exactly captures the joint distribution of nodes which end up sharing. The existence of such distributions has advantages in simulating and estimating the performance of the system. We establish this result via a general theorem characterizing which types of models lead to coverage processes, and also show that all coverage processes possess the desirable submodular properties. We complement our theoretical results with experiments on several real-world P2P topologies. We compare our model quantitatively against more naïve models ignoring network effects. A main outcome of the experiments is that a good incentive scheme should make the reward dependent on a node’s degree in the network.

## 1 Introduction

Peer-to-Peer (P2P) file sharing systems have become an important platform for the dissemination of files, music, and other content. The basic idea is very simple: individuals make files available for download from their own machine. Other users can search for files they desire and download them from a peer who has made the file available. Naturally, designing systems such that the search and download of files are efficient poses many research challenges, which have received a lot of attention in the literature [2, 22].

A second, and somewhat orthogonal, issue is how to ensure sufficient participation and sharing of files. Unless enough content is provided by individuals, the utility of membership will be very small. If free-riding [9] is too prevalent, the system may exhibit a quick decrease in membership common to public-goods type economic settings [23].

Thus, the P2P system must be designed with incentives in mind to encourage file sharing. These incentives can take the form of monetary payments or redeemable “points” [11], download privileges, or simply recognition. From the system designer’s perspective, these payments should be “small,” while ensuring enough participation.

---

\*Department of Computer Science, University of Southern California, CA 90089-0781, USA. E-mail: salek@usc.edu

<sup>†</sup>Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052-6399, USA. E-mail: shahins@microsoft.com. Work done while the author was at the University of Southern California.

<sup>‡</sup>Department of Computer Science, University of Southern California, CA 90089-0781, USA. E-mail: dkempe@usc.edu. Work supported in part by NSF CAREER Award 0545855 and an ONR Young Investigator Award.

On the other hand, from a peer’s perspective, the payments need to be weighed against the cost incurred by sharing a file. In this paper, we assume that the content is shared legally and the system is designed with security in mind: hence, the main cost to an individual is the upload bandwidth which will be used whenever another peer downloads a file from this node.

Nodes will in general choose to download from nearby peers (in terms of bandwidth or latency). Therefore, as additional nearby peers share the same files, the load will get distributed among more nodes, and the cost to each individual node will decrease. Thus, not only will we expect cascading effects of sharing based on social dynamics [12], but we would also expect these cascading effects to be based on a network structure determined by point-to-point latencies and bandwidths.

Our contribution in this paper is the definition and analysis (both theoretical and experimental) of a natural model for peers’ sharing behavior in P2P systems, in the presence of network effects and economic incentives. In our model, we focus only on sharing one file; in practice, the model can be applied separately for each file of interest. The basic premise of the model is that each node has a certain *demand* for the file. Furthermore, the network determines which percentage of the demand will be met by downloading from each peer sharing the file<sup>1</sup>. The crucial implication of this model is that the more nearby peers are sharing a file, the more evenly the demand will be distributed among them.

The upload bandwidth cost is compensated by a *payment* to the peers who make the file available. Again, our model is agnostic about whether these payments are monetary, recognition, or take other forms. In our model, the payments can be explicitly based on the network degree of peers, since high-degree nodes presumably serve a key role in propagating sharing behavior.

We argue that this model captures the essential dynamics of P2P systems in which a peer can join the network and download files without sharing; hence, availability of files is not the only incentive for sharing. The FastTrack P2P protocol, used by KaZaA, Grokster, and iMesh, is an example where this assumption holds; hence, our model should be a reasonable approximation for these services in terms of its incentives.

The network operator is interested in maximizing a social welfare function  $W$ , which grows monotonically as a function of the set of nodes that share the file. This function could be the total number of sharing nodes, the number of nodes with at least one uploading neighbor, or the total download bandwidth available to peers under various natural models of downloading.

After defining this model formally (in Section 2), we prove strong and general *diminishing returns* properties about it (in Section 3). In particular, we show that whenever  $W$  is monotone and submodular, the network’s social welfare as a function of the payments offered to the peers is monotone i.e., increasing payments will always increase social welfare. However the *rate* of increase decreases when payments are already high. We call the latter property diminishing returns.

To prove this result, we consider a slightly different model, wherein payments are combined with giving the network operator the ability to “force” some set  $S$  of peers to share. By first proving certain local submodularity properties for this modified model, the desired diminishing returns properties are implied by the general result of Mossel and Roch [18]. However, we derive a similar result to [18] for a broad subclass of submodular functions which we call coverage functions. It consists of the functions for which in the underlying process, the distribution of nodes sharing the file is equivalent to the distribution of nodes reachable from  $S$  in an appropriately defined random graph model. We establish this equivalence via a general and non-trivial theorem characterizing all functions that can be obtained by counting reachable nodes under random graph models. As a corollary, our approach provides a much simpler proof of the main result from [18] for coverage processes. Moreover, the fact that the propagation of sharing behavior is a coverage process is useful for the purpose of simulating the process and estimating the parameters of the system, allowing more efficient algorithms for simulations. Finally, our characterization can be of independent interest in the study of submodular set functions.

While the bulk of our paper focuses on a theoretical analysis of the demand model, we complement the theoretical results by an experimental evaluation of our model (in Section 4), using two network topologies derived from real-world data sets [13, 20, 21], and a regular two-dimensional grid topology. We first show

---

<sup>1</sup>In practice, we could expect these percentages to correlate strongly with network latency or available bandwidth, but our model is agnostic about the derivation.

that network effects are significant by comparing our demand model with one in which peers are not aware of changes in load due to nearby sharing peers. We then evaluate different payment schemes, in particular regarding their dependence on nodes' degrees. We evaluate these both in terms of the fraction of peers that end up sharing, and the amount paid by the network operator per sharing node.

## 1.1 Related Work

There is a large body of work on incentive mechanisms in P2P file-sharing systems. (See [8] for a thorough overview and [27] for a recent generalized analysis framework.) Incentive mechanisms can be classified in three categories: barter-based mechanisms, reputation-based mechanisms, and currency-based mechanisms.

*Barter-based* methods [1] enforce repeated transactions among peers by matching each peer to only a small subset of the network, hence raising the survival chance for strategies based on reciprocity. This method only works when we have a small and popular set of files. For instance, the BitTorrent protocol [6] is a popular P2P file-sharing protocol using this method.

*Reputation-based mechanisms* have an excellent track record at facilitating cooperation in very diverse settings, from evolutionary biology to marketplaces like eBay. These systems keep a tally of the contribution of each peer; the past contributions determine which peers obtain more of the system's resources in the future. However, the availability of cheap pseudonyms in P2P systems makes reputation systems vulnerable to Sybil and whitewashing attacks [9], leading to ongoing work on designing sybilproof reputation mechanisms [5]. Moreover, reputation systems may be vulnerable to coordinated gaming strategies due to distributed rating systems [24].

Inspired by markets, a P2P system can also deploy a currency scheme to facilitate resource contributions by rational peers. Generally, peers earn currency by contributing resources to the system, and spend the currency to obtain resources from the system. Karma [25] is one example of this kind. *Currency-based systems* may also suffer from Sybil and whitewashing attacks, depending on their policies toward newcomers. If newcomers are endowed with a positive balance, then the system is vulnerable to these attacks; otherwise, there might not be enough incentive for newcomers to join the network. Balance control could also be troublesome, as the system might need to deal with negative balances.

Lai et al. [16] introduced the concept of "private" history vs. "shared" history as a way to combine barter-based and reputation-based mechanisms in the context of an evolutionary prisoner's dilemma. Shared history is a pool that records peers' past behavior and services them according to their reputation. In [9], file sharing is modeled as a social phenomenon, akin to those discussed by Schelling [23]. Users consider whether or not to contribute files based on the number of other users who contribute. Our model is different in that it explicitly models the costs incurred by contributing nodes, rather than simply positing an intrinsic generosity parameter for each user.

## 2 Models and Preliminaries

We consider a peer-to-peer network with  $n$  servers (or *nodes* or *peers*), and focus on the behavior of sharing one particular file. Thus, each peer  $v$  may either choose to share the file or to not share it. We also call sharing peers *active*, and the other ones *inactive*. The set of all peers who share is denoted by  $V^+$ .

### 2.1 The Demand Model

Each peer has a local *demand*  $d_v$  for the file: this demand will originate from individual users on the server  $v$  (who themselves might not possess the file or be in a position to make it available). The demand  $d_v$  should be served by downloading the file from other servers  $u \in V^+$ . The *quality* of the connection between  $v$  and  $u$  is captured by a matrix  $P$ : the larger  $p_{v,u}$ , the larger a fraction of  $v$ 's demand will be served by  $u$  (assuming that  $u$  shares the file). Specifically, the demand that  $u \in V^+$  will see from  $v$  is  $d_v \cdot \frac{p_{v,u}}{\sum_{w \in V^+} p_{v,w}}$ . The matrix  $P$  will in practice depend on network latencies or bandwidth, as well as explicit download agreements. It need not be symmetric. For the purpose of the general model, we are agnostic to the derivation of  $P$ ; in Section

4, we will derive  $P$  from measured network latencies by positing a latency threshold which individuals are willing to tolerate.

A node  $u \in V^+$  sharing the file will incur a *cost* of  $c_u$  per unit of demand that it serves; this cost is the result of using upload bandwidth, machine processing time, or similar resources. To encourage peers to share the file despite this cost, the P2P network administrator offers payments  $\pi_u$  to the nodes  $u \in V^+$ . These payments need not be the same for all nodes, and can be derived from the network structure, e.g., a node's degree.

Different nodes may have different (and unknown) tradeoffs between money and upload bandwidth. We model this fact by assuming that each node  $u$  has a tradeoff factor  $\lambda_u$ , drawn independently and uniformly at random from  $[0, 1]$ , which captures how many units of bandwidth one unit of money is worth to the node. Thus, the *sharing utility* of an active node  $u \in V^+$  is

$$U(u) = \lambda_u \pi_u - c_u \sum_v \frac{d_v p_{v,u}}{\sum_{w \in V^+} p_{v,w}},$$

while the sharing utility of non-sharing nodes is 0. (A non-sharing node does not get paid and incurs no upload costs.) We assume that agents are rational, and thus choose whether to share or not to share so as to maximize their own utility.

## 2.2 Other Models

As we discussed in Section 1, one of our main contributions is the observation that file sharing behavior should be subject to positive network externalities, i.e., that the presence of other sharing peers makes sharing less costly. To quantify the size of such network effects, we define two alternative models with no or limited effects; we will compare these two models experimentally with the demand model in Section 4.

1. In the *No-Network* Model, the peers completely ignore other sharing peers. Thus, a node  $u$  assumes that if it shares the file, then it will see a fraction  $p_{v,u}$  of the demand originating with node  $u$ . Hence, the perceived utility of node  $u$  when sharing is

$$U(u) = \lambda_u \pi_u - c_u \cdot \sum_v d_v p_{v,u}.$$

2. In the *One-Hop* Model, the peers are aware of network effects in a very limited way: node  $u$  assumes that any node  $v$  sharing the file will contribute toward serving both  $v$ 's and  $u$ 's demand, but not toward serving the demand of any other node  $w \neq u, v$ . Thus, the perceived utility of node  $u \in V^+$  is in the *One-Hop* Model is

$$U(u) = \lambda_u \pi_u - c_u \cdot \frac{d_u p_{u,u}}{\sum_{w \in V^+} p_{u,w}} - c_u \cdot \sum_{v \neq u} \frac{d_v p_{v,u}}{\sum_{w \in V^+ \cap \{u,v\}} p_{v,w}}.$$

## 2.3 Payment Schemes, Sharing Process, and Administrator's Objective

The network administrator's choice is how to set the payment offers  $\pi_u$ . In doing so, the administrator balances two competing goals: low overall payments and high utility for the participants in the system. In this paper, we study the impact of payment schemes on these objectives.

In order to provide enough incentives for sharing, the network administrator should always ensure that  $\pi_u \geq C_u := c_u \cdot \sum_v d_v$ . Otherwise, even a node  $u$  with  $\lambda_u = 1$  (i.e., the highest possible utility for money) would have no incentive to share the file if no other peers are sharing the file.

The full model is thus as follows: after the administrator decides on the payments  $\pi_u$  for all nodes  $u$ , the random tradeoffs  $\lambda_u$  between money and bandwidth are determined independently for all nodes  $u$ . Subsequently, the process proceeds in iterations. In each iteration, all peers simultaneously decide whether to share the file or not, based on the payments, costs, and previous decisions of all other peers. The process

continues until an equilibrium is reached. Notice that because the cost to a peer is monotone decreasing in the set  $V^+$  of currently sharing peers, the set of sharing peers can only become larger from iteration to iteration. In particular, this implies that the process will eventually terminate with some set  $V^+$  of active peers. We call this the *sharing process* or *activation process*.

The network administrator is in general interested in increasing access to the file while keeping the payments low. This general objective may be captured using various metrics. In general, we allow for any overall social welfare function  $W$  which increases monotonically in the set  $S$  of sharing nodes. Notice that since the set  $S$  itself is the result of a random process, the administrator's goal will be to maximize  $E[W(S)]$ , where  $S$  is derived from the random activation process in the demand model. Several social welfare functions  $W$  suggest themselves naturally:

1. The number of active peers is a natural measure of participation. It is the measure frequently studied in the context of the diffusion of innovations or behaviors in social networks [10,12,14,15,17,18]. While the objective is similar, the precise dynamics are different between those models and the demand model.
2. The total number of *served* nodes, i.e., nodes  $v$  with at least one active node  $u$  with  $p_{u,v} > 0$ . This model is appropriate if we only care about how *many* peers can download the file, but not about the quality of the connection. It implicitly assumes that each peer has a constant utility of 1 for downloading.
3. Each node  $u$  gets a utility of  $\sum_{v \in V^+} p_{u,v}$ , and the social welfare is the sum of all these utilities. This model is based on the assumption that  $u$ 's demand is served by all of its neighbors (including possibly  $u$ ) simultaneously, and that  $u$ 's utility is the total "download bandwidth" available in this sense. We call this the *sum-welfare* function.
4. Each node  $u$  gets a utility of  $\max_{v \in V^+} p_{u,v}$ , and the social welfare is the sum of all these utilities. This is based on the assumption that  $u$ 's demand is served by its active neighbor with the best connection, corresponding to a situation where parallel download from multiple sources is not possible. We call this the *max-welfare* function.

Notice that the social welfare function  $W$  may also include the utilities of the sharing nodes.

### 3 Theoretical Analysis of the Model

The main analytical contribution of this paper is based on *coverage processes*<sup>2</sup>, defined formally in Definition 5. Informally, a coverage process is a random process such that the distribution over sets of ultimately active nodes is also the distribution of reachable nodes under a suitably chosen distribution of random graphs. Our results on coverage processes are twofold:

(1) We give a general characterization of coverage processes, and show that the activation process for P2P systems is a coverage process. (2) We give a significantly simplified proof (compared to the general result of [18]) showing that under coverage processes, the expected social welfare as a function of the payments has diminishing returns in the sense of Definition 1 so long as the social welfare is a submodular function of the active nodes.

Recall that a function  $f$  defined on sets is *submodular* if  $f(S + v) - f(S) \geq f(T + v) - f(T)$  whenever  $S \subseteq T$ , i.e., if the addition of an element to a larger set causes a smaller increase in the function value than to a smaller set. Thus, submodularity is the discrete analogue of concavity, and intuitively corresponds to "diminishing returns." An easy inductive proof (on the size of  $X$ ) shows that submodularity is equivalent to the condition that for all sets  $X$ ,

$$f(S \cup X) - f(S) \geq (T \cup X) - f(T) \quad \text{whenever } S \subseteq T. \quad (1)$$

---

<sup>2</sup>We thank Bobby Kleinberg for this naming suggestion, and also note here that Theorem 8 was derived independently by him.

**Definition 1** A function  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  has diminishing returns if for every pair  $i, j$  and all vectors  $\mathbf{x}$ , it satisfies

$$\frac{\partial g(x_1, x_2, \dots, x_n)}{\partial x_i \partial x_j} \leq 0.$$

**Remark 2** The notion of “diminishing returns” is strictly weaker than concavity; it corresponds to concavity only along positive coordinates axes.<sup>3</sup>

The two main contributions of our paper together imply the following theorem as a corollary:

**Theorem 3** Let  $\bar{W}(\pi_1, \dots, \pi_n) = \mathbb{E}[W(S)]$  be the expected social welfare when set  $S$  is obtained from the sharing process of the demand model with payments  $\pi_1, \dots, \pi_n$ .

If  $W(S)$  is submodular, then  $\bar{W}(\pi_1, \dots, \pi_n)$  is monotone and has diminishing returns with respect to the payments  $\pi_1, \dots, \pi_n$ .

For the social welfare function, the diminishing returns property intuitively means that the additional benefit in social welfare that can be derived from increasing the payment to a peer  $u$  decreases as the peers’ current payments increase.

The proof of Theorem 3 is based on analyzing the following *Seed Set Model*, which we define mainly for the purpose of analysis.

**Definition 4 (Seed Set Model)** For each node, the payment offered is  $\pi_u = C_u$ . Besides payments, we have a seed set  $S$  of peers that will always share regardless of the payments. Subsequently, the process unfolds exactly according to the sharing process.

The main technical step is to show that the Seed Set Model is a *coverage process*, in the following sense.

**Definition 5 (Coverage Process)** Let  $\phi(S)$  be the random variable describing the set of nodes active at the end of a process starting from the set  $S$  of nodes active. The process is called a coverage process if there exists a distribution  $D$  over graphs  $G$  such that for each set  $T$  of nodes,  $\text{Prob}[\phi(S) = T]$  equals the probability that exactly  $T$  is reachable starting from  $S$  in  $G$  if  $G$  is drawn from the distribution  $D$ .

**Remark 6** Without using our nomenclature, [14] showed submodularity for the Cascade and Threshold models of innovation diffusion [10,12] by establishing that both gave rise to coverage processes. Subsequently, [15] showed that there are natural diffusion processes which are not coverage processes, yet have a submodular function  $\mathbb{E}[|\phi(S)|]$ .

We prove that the Seed Set Model is a coverage process in two steps. First, in Section 3.1, we give a general and complete characterization of Coverage Processes. This characterization may be of interest in its own right, as coverage processes have a practical advantage: they can be simulated easily and efficiently, by first generating a random graph according to  $D$ , and then simply finding the set of reachable nodes.

Then, in Section 3.2, we show that the Seed Set Process satisfies the conditions established in Section 3.1. Finally, in Section 3.3, we give a simple proof that for *any* coverage process and any submodular social welfare function, the expected social welfare under the process is also submodular. This implies diminishing returns with respect to the payments.

**Remark 7** The fact that the tradeoffs  $\lambda_u$  between money and bandwidth are uniformly random in  $[0, 1]$  is important to ensure the submodularity and diminishing returns properties. If the  $\lambda_u$  are not random but fixed, then the diminishing returns and submodularity properties cease to hold. Furthermore, in the Seed Set Model, the optimization problem of finding the best seed set  $S$  of at most  $k$  nodes becomes very hard, as we show in the appendix.

---

<sup>3</sup>We thank Shaddin Dughmi for pointing out this ambiguity in an earlier version of the paper.



### 3.1 Characterization of Coverage Processes

In this section, we characterize *exactly* which random processes are coverage processes. This theorem may be of interest in its own right, when analyzing different processes.

Our setting is exactly as in the paper by Mossel and Roch [18]: each node  $u$  has an activation function  $f^u$ , which is monotone non-decreasing and satisfies  $f^u(\emptyset) = 0$ . Each node independently chooses a threshold  $\theta_u \in [0, 1]$  uniformly at random, and becomes active when  $f^u(S) \geq \theta_u$ , where  $S$  is the previously active set of nodes. This process is repeated until no more changes occur.

In order to express our results concisely, we use the following discrete equivalent of a derivative (see, e.g., [26]). For a function  $f$  defined on sets, we define inductively:

$$\begin{aligned} f_{\emptyset}(S) &= f(S) \\ f_{R \cup \{v\}}(S) &= f_R(S \cup \{v\}) - f_R(S). \end{aligned}$$

It is not difficult to verify that this notion is well-defined, i.e., independent of *which* element  $v$  is chosen at which stage.

**Theorem 8** *The following conditions are necessary and sufficient for the process to be a coverage process.*

- For all sets  $T$  of odd cardinality  $|T|$ , as well as for  $T = \emptyset$ , and each node  $u$ , we have  $f_T^u(\overline{T}) \geq 0$ .
- For all sets  $T$  of positive even cardinality  $|T|$ , and each node  $u$ , we have  $f_T^u(\overline{T}) \leq 0$ .
- $f^u(\emptyset) = 0$  for all  $u$ .

To prove this theorem, we begin with the following reasoning. Focus on one node  $u$ , and its activation function  $f^u$ . If there were an equivalent graph distribution  $D$ , then it would have to define a probability  $q_u(T)$  for the presence of edges from exactly the vertex set  $T$  to  $u$ . These probabilities need to satisfy the following property: if a set  $S$  of nodes is active, then the probability of  $u$  having *at least* one incoming edge from  $S$  must equal  $f^u(S)$ . Thus, a necessary and sufficient condition for being a coverage function is that for each node  $u$ , there exists a distribution  $q_u(T)$  over sets  $T$  such that

$$f^u(S) = \sum_{T: T \cap S \neq \emptyset} q_u(T). \quad (2)$$

We can express this requirement more compactly using matrix notation. Let  $\mathbf{f}_u$  be the  $(2^n - 1)$ -dimensional vector consisting of all entries of  $f^u(S)$  for  $S \neq \emptyset$ . Similarly, let  $\mathbf{q}_u$  be the  $(2^n - 1)$ -dimensional vector of all  $q_u(S)$  for  $S \neq \emptyset$ . Let  $A$  be the  $((2^n - 1) \times (2^n - 1))$ -dimensional matrix indexed by non-empty subsets such that  $A_{S,T} = 1$  if and only if  $S \cap T \neq \emptyset$ , and  $A_{S,T} = 0$  otherwise. ( $A$  is called an *incidence matrix* [4].) Then, Equation 2 can be rewritten as the requirement that for each node  $u$ , there exists a distribution  $\mathbf{q}_u$  such that  $A \cdot \mathbf{q}_u = \mathbf{f}_u$ .

For the analysis, we fix a canonical ordering of subsets. Specifically, if the current (sub-)universe consists of  $k$  nodes indexed  $\{1, 2, \dots, k\}$ , their canonical ordering is defined recursively as first containing all subsets of  $\{1, 2, \dots, k-1\}$  in canonical order, then the set  $\{k\}$ , followed by the sets  $T \cup \{k\}$ , where the sets  $T \subseteq \{1, 2, \dots, k-1\}$  appear in canonical order.

In order to find out when the distribution  $\mathbf{q}_u$  exists, we want to solve the equation  $A \cdot \mathbf{q}_u = \mathbf{f}_u$ , or  $\mathbf{q}_u = A^{-1} \cdot \mathbf{f}_u$ . While the inverses of some incidence matrices have been studied before (see, e.g., [3]), we are not aware of any source explicitly giving the inverse of the matrix  $A$ . Hence, we establish here:

**Lemma 9** *The inverse of  $A$  is the matrix  $B$  defined by*

$$b_{S,T} := \begin{cases} 0 & \text{if } S \cup T \neq \{1, \dots, n\} \\ (-1)^{|S \cap T|+1} & \text{otherwise} \end{cases}$$

**Proof.** The key insight is that under the canonical ordering of sets defined above, the matrices  $A$  and  $B$  can be defined recursively via matrices  $A_k$  and  $B_k$ . Specifically, let  $A_1 = 1$ , and

$$A_{k+1} = \left( \begin{array}{c|c|c} & 0 & \\ & \vdots & A_k \\ & 0 & \\ \hline 0 \dots 0 & 0 & 1 \dots 1 \\ \hline & 1 & \\ A_k & \vdots & 1 \\ & 1 & \\ & 1 & \end{array} \right).$$

Similarly, let  $B_1 = 1$ , and

$$B_{k+1} = \left( \begin{array}{c|c|c} & 0 & \\ & \vdots & B_k \\ & 0 & \\ & -1 & \\ \hline 0 \dots 0 & 0 & 1 \\ \hline & 0 & \\ B_k & \vdots & -B_k \\ & 0 & \\ & 1 & \end{array} \right).$$

The fact that  $A = A_n$  and  $B = B_n$  can be observed directly from the definition and the canonical ordering.

To prove the lemma, we can show by induction on  $k$  that  $A_k \cdot B_k = I_k$  for all  $k$ , where  $I_k$  is the  $k \times k$  identity matrix. The base case  $k = 1$  is obvious. For the inductive step to  $k + 1$ , consider the  $(i, j)$  entry  $(A_{k+1} \cdot B_{k+1})_{i,j}$ . We distinguish 7 different cases, based on the  $(i, j)$  indices. (We use 0 to denote the  $(2^k - 1) \cdot (2^k - 1)$  matrix of all zeroes,  $\mathbf{1}$  for the vector of all ones, and  $\hat{\mathbf{u}}$  for the  $(2^k - 1)$ -dimensional unit vector with 1 in its last coordinate and 0 everywhere else.)

1. If  $i, j < 2^k$ , then the entry is  $(A_k \cdot 0)_{i,j} + 0 + (A_k \cdot B_k)_{i,j} = (I_k)_{i,j}$  by induction hypothesis.
2. If  $i > 2^k, j < 2^k$ , then (writing  $i' = i - 2^k$ ), the entry is  $(A_k \cdot 0)_{i',j} - \hat{\mathbf{u}}_{i'} + (\mathbf{1} \cdot B_k)_{i'} = 0$  using Lemma 10(a) below.
3. If  $i < 2^k, j > 2^k$ , then (writing  $j' = j - 2^k$ ), the entry is  $(A_k \cdot B_k)_{i,j'} + 0_{i,j'} - (A_k \cdot B_k)_{i,j'} = 0$ .
4. If  $i, j > 2^k$ , then (writing  $i' = i - 2^k, j' = j - 2^k$ ), the entry is  $(A_k \cdot B_k)_{i',j'} + \hat{\mathbf{u}}_{j'} - (\mathbf{1} \cdot B_k)_{j'} = (I_k)_{i',j'}$ , again using Lemma 10(a).
5. If  $i = j = 2^k$ , a straightforward calculation shows that the entry is 1.
6. If  $i = 2^k, j < 2^k$ , then the entry is  $(\mathbf{1} \cdot B_k)_j - \hat{\mathbf{u}}_j = 0$  by Lemma 10(a). Similarly, for  $i = 2^k, j > 2^k$ , writing  $j' = j - 2^k$ , the entry is  $\hat{\mathbf{u}}_{j'} - (\mathbf{1} \cdot B_k)_{j'} = 0$  by Lemma 10(a).
7. Finally, for  $j = 2^k, i < 2^k$ , the entry is  $-(A_k \cdot \hat{\mathbf{u}}^T)_i + (A_k \cdot \hat{\mathbf{u}}^T)_i = 0$ , whereas for  $j = 2^k, i > 2^k$ , writing  $i' = i - 2^k$ , the entry is  $-(A_k \cdot \hat{\mathbf{u}}^T)_{i'} + 1 = 0$  by Lemma 10(b).

This proves that  $A_{k+1} \cdot B_{k+1} = I_{k+1}$ . ■

**Lemma 10** Let  $\mathbf{1}$  be the vector of all 1's, and  $\hat{\mathbf{u}}$  defined as in the proof of Lemma 9. Then, (a)  $\mathbf{1} \cdot B_k = \hat{\mathbf{u}}$ , and (b)  $A_k \cdot \hat{\mathbf{u}}^T = \mathbf{1}$ .



**Proof.** For part (a), we show that the row sums of all rows of  $B_k$  are zero except the last row, which has a row sum of one. The proof is by induction. The base case  $B_1 = 1$  is clear. For the inductive step from  $k$  to  $k+1$ , first notice that all the entries in columns  $j < 2^k - 1$  are zero by induction hypothesis. For column  $2^k - 1$ , the row sum of  $B_k$  contributes 1 by induction hypothesis, from which 1 is subtracted because of the entry in the middle column. Column  $2^k$  adds up to 0 explicitly, and columns  $j = 2^k + 1, \dots, 2^{k+1} - 2$  have terms of  $B_k$  and  $-B_k$  canceling out. Finally, for the last column, the entries of  $B_k$  and  $-B_k$  cancel out, leaving the entry 1 from the middle column.

For part (b), simply notice that using part (a) and the induction hypothesis of Lemma 9 (for  $k$ ), we get that  $A_k \cdot \hat{\mathbf{u}}^T = A_k \cdot B_k^T \cdot \mathbf{1} = I_k \cdot \mathbf{1} = \mathbf{1}$ . Here, we used that  $B_k$  is symmetric. ■

The next lemma shows that so long as all  $q_u(S)$  are non-negative, by setting  $q_u(\emptyset)$  appropriately, we can always obtain a probability distribution.

**Lemma 11** *With  $q_u(S)$  defined as  $\mathbf{q}_u = B \cdot \mathbf{f}_u$ , we have  $\sum_S q_u(S) \leq 1$ .*

**Proof.** Let  $\mathbf{1}$  denote the all-ones vector as before. We can rewrite

$$\sum_S q_u(S) = \mathbf{1} \cdot (B \cdot \mathbf{f}_u) = (\mathbf{1} \cdot B) \cdot \mathbf{f}_u.$$

Using Lemma 10(a), the sum is exactly equal to  $f^u(\{1, \dots, n\}) \leq 1$ , completing the proof. ■

By Lemma 9, we know that  $\mathbf{q}_u = B \cdot \mathbf{f}_u$ . And by Lemma 11, the entries sum up to at most 1. Thus, it remains to show that the entries of  $\mathbf{q}_u$  are non-negative if and only if  $f^u$  satisfies the conditions of Theorem 8. To relate these formulations, we prove the following non-recursive characterization of discrete derivatives.

**Lemma 12** *For all sets  $T$ , we have that*

$$f_T(W) = \sum_{S \subseteq T} (-1)^{|T|-|S|} f(W \cup S).$$

**Proof.** The proof is by induction on  $|T|$ . For  $T = \emptyset$ , the claim is trivial. Now, consider a set  $T_{k+1} = T_k \cup \{t\}$  of size  $k+1$ . By definition of the discrete derivative and induction hypothesis,

$$\begin{aligned} f_{T_{k+1}}(W) &= f_{T_k}(W \cup \{t\}) - f_{T_k}(W) \\ &= \sum_{S \subseteq T_k} (-1)^{k-|S|} f(W \cup S \cup \{t\}) - \sum_{S \subseteq T_k} (-1)^{k-|S|} f(W \cup S) \\ &= \sum_{S \subseteq T_{k+1}: t \in S} (-1)^{k+1-|S|} f(W \cup S) + \sum_{S \subseteq T_{k+1}: t \notin S} (-1)^{k+1-|S|} f(W \cup S) \\ &= \sum_{S \subseteq T_{k+1}} (-1)^{k+1-|S|} f(W \cup S), \end{aligned}$$

which completes the inductive proof. ■

**Proof of Theorem 8.** Fix any node  $u$ , and define  $\mathbf{q}_u = B \cdot \mathbf{f}_u$ . By Lemma 12, we can write the discrete derivative of  $f^u$  at  $\bar{T}$  as

$$f_T^u(\bar{T}) = \sum_{S \subseteq T} (-1)^{|T|-|S|} f^u(\bar{T} \cup S).$$

Now, if  $|T|$  is odd, then  $(-1)^{|T|-|S|} = (-1)^{|S|+1}$ , so we can rewrite the above as

$$\sum_{S \subseteq T} (-1)^{|S|+1} f^u(\bar{T} \cup S) = \sum_{W \supseteq \bar{T}} (-1)^{|W \cap T|+1} f^u(W) = q_u(T).$$

Similarly, if  $|T|$  is even, then  $(-1)^{|T|-|S|} = (-1)^{|S|}$ , so we can rewrite the discrete derivative as

$$\sum_{S \subseteq T} (-1)^{|S|} f^u(\bar{T} \cup S) = \sum_{W \supseteq \bar{T}} (-1)^{|W \cap T|} f^u(W) = -q_u(T).$$

Thus, the  $q_u(T)$  are all non-negative (and the probability distribution thus well-defined) if and only if  $f_T^u(\bar{T}) \geq 0$  for  $|T|$  odd, and  $f_T^u(\bar{T}) \leq 0$  for  $|T| > 0$  even.  $\blacksquare$

### 3.2 Coverage Property of the Seed Set Process

In this section, we establish the following theorem.

**Theorem 13** *The Seed Set Process is a coverage process.*

**Proof.** In order to prove this theorem, we want to apply Theorem 8. To do so, we need to show that the local decisions of nodes about sharing can be cast in terms of submodular threshold functions. Specifically, we define

$$f^u(S) := 1 - \frac{1}{C_u} \cdot c_u \cdot \sum_v \frac{d_v p_{v,u}}{\sum_{w \in S \cup \{u\}} p_{v,w}}$$

and let  $\theta_u = 1 - \frac{\lambda_u \pi_u}{C_u}$ . (Recall from Section 2.3 that  $C_u = c_u \cdot \sum_v d_v$ .)

A node  $u$  becomes active if doing so has positive utility, i.e., if  $\lambda_u \pi_u > c_u \cdot \sum_v \frac{d_v p_{v,u}}{\sum_{w \in S \cup \{u\}} p_{v,w}}$ . Dividing both sides by  $C_u$ , and subtracting from 1 shows that this is equivalent to saying that

$$1 - \frac{\lambda_u \pi_u}{C_u} < 1 - \frac{1}{C_u} \cdot c_u \cdot \sum_v \frac{d_v p_{v,u}}{\sum_{w \in S \cup \{u\}} p_{v,w}}.$$

Since  $\lambda_u \pi_u$  is uniformly random in  $[0, C_u]$  by the definition of  $\pi_u$  in the Seed Set Model, this condition is equivalent to saying that  $\theta_u < f^u(S)$ . Thus, we have shown that the activation process can be equivalently recast in terms of threshold activations functions.

Finally, we need to show that for every node  $u$ , all derivatives  $f_T^u(S)$  are non-negative when  $|T|$  is odd and non-positive when  $|T| > 0$  is even. (The fact that  $f^u(S) = f_\emptyset^u(S)$  is non-negative follows directly by definition.) Let

$$\hat{f}^u(x_1, \dots, x_n) = 1 - \frac{1}{C_u} \cdot c_u \cdot \sum_v \frac{d_v p_{v,u}}{\sum_{v_i \in V} p_{v,v_i} x_i}$$

be the continuous equivalent of the local influence function  $f^u$ . For a set  $S$ , let  $\mathbf{y}^{(S)}$  denote the  $n$ -dimensional vector with  $y_i^{(S)} = 1$  if  $v_i \in S \cup \{u\}$  and  $y_i^{(S)} = 0$  otherwise. Then,  $f^u(S) = \hat{f}^u(\mathbf{y}^{(S)})$ . Notice that by definition, there is no division by zero.

Writing  $dY_T = dy_{i_1} dy_{i_2} \dots dy_{i_{|T|}}$ , where  $T = \{i_1, i_2, \dots, i_{|T|}\}$ , an easy inductive proof first shows that

$$f_T^u(S) = \int_0^1 \dots \int_0^1 \frac{d\hat{f}^u(\mathbf{y}^{(S)})}{dY_T} dY_T.$$

It remains to show that each term inside the integration is non-negative for odd  $|T|$  and non-positive for even  $|T|$ . We accomplish this by showing that

$$\frac{d\hat{f}^u(\mathbf{y}^{(S)})}{dY_T} = (-1)^{|T|+1} |T|! \frac{c_u}{C_u} \sum_v \frac{d_v p_{v,u} \prod_{t \in T} p_{v,t}}{(\sum_{v_i \in V} p_{v,v_i} y_i^{(S)})^{|T|+1}}.$$

The proof is by induction. The base case:  $|T| = 1$  can be verified easily. Assume that the claim holds for  $|T| = i - 1$ . We have

$$\begin{aligned}
\frac{d\hat{f}^u(\mathbf{y}^{(S)})}{dY_T dy_i} &= \frac{d}{dy_i} (-1)^{|T|+1} |T|! \frac{c_u}{C_u} \sum_v \frac{d_v p_{v,u} \prod_{t \in T} p_{v,t}}{(\sum_{v_i \in V} p_{v,v_i} y_i)^{|T|+1}} \\
&= (-1)(-1)^{|T|+1} |T|! \frac{c_u}{C_u} \cdot \sum_v \frac{(|T|+1) p_{v,v_i} d_v p_{v,u} \prod_{t \in T} p_{v,t} (\sum_{v_i \in V} p_{v,v_i} y_i)^{|T|}}{(\sum_{v_i \in V} p_{v,v_i} y_i)^{2|T|+2}} \\
&= (-1)^{|T|+2} |T+1|! \frac{c_u}{C_u} \sum_v \frac{d_v p_{v,u} \prod_{t \in T \cup \{v_i\}} p_{v,t}}{(\sum_{v_i \in V} p_{v,v_i} y_i)^{|T|+2}}.
\end{aligned}$$

This completes the inductive proof, and thus the proof of Theorem 13.  $\blacksquare$

While we defined the Seed Set Process primarily as a tool for analysis, we remark here that Theorem 13 has a direct consequence for the optimization problem of maximizing the expected total number of active nodes at the end of the process, subject to a size constraint on the seed set  $S$ . A Theorem of Nemhauser et al. [7, 19] states that if  $f$  is any non-negative, monotone, and submodular function on sets, then the greedy algorithm is a polynomial-time  $(1 - 1/e)$ -approximation (where  $e$  is the base of the natural logarithm). Since we can approximate the expected number of active nodes under the Seed Set Process arbitrarily closely by simulating the activation process (see [14] for an in-depth discussion of the greedy algorithm), we obtain the following corollary:

**Corollary 14** *The best starting set  $S$  for the Seed Set Process can be approximated within  $(1 - 1/e - \epsilon)$  in polynomial time, for any  $\epsilon > 0$ .*

### 3.3 Diminishing Returns of Expected Social Welfare

Finally, we use the machinery of coverage processes to show diminishing returns of social welfare. Consider an arbitrary coverage process. When the coverage process starts with the set  $T$ , let  $\phi(T)$  be a random variable describing the set of nodes active at the end of the process. Thus, the distribution of  $\phi(T)$  for all  $T$  precisely characterizes the coverage process. Our main theorem is now the following:

**Theorem 15** *Let  $h(S)$  be any monotone submodular function of  $S$ . Then,  $\mathbb{E}[h(\phi(T))]$  is a monotone submodular function of  $T$ , where the expectation is taken over the randomness in  $\phi(T)$ .*

This theorem follows from the general result of [18], since all coverage processes are locally submodular, and our utility function is submodular with respect to the set of sharing neighbors. However, below we give a very simple proof based on reachability in graphs using the fact that  $\phi$  is a coverage process. This is useful for the purpose of simulating the process and estimating  $\phi$ . It means that instead of generating random thresholds and simulating a dynamic process, we can generate a random graph and then simply use BFS to find the number of reachable nodes.

**Proof.** Because  $\phi$  is a coverage process, by Theorem 8, there is a distribution  $\Pr[\cdot]$  over graphs  $H$  such that for any set  $T$ , the set of nodes reachable in  $H$  from  $T$  has the same distribution as  $\phi(T)$ . Let  $\phi_H(T)$  denote the set of nodes reachable from  $T$  in  $H$ . Then,

$$\mathbb{E}[h(\phi(T))] = \sum_H \Pr[H] \cdot h(\phi_H(T)).$$

Fix some graph  $H$  and let  $S \subseteq T$  and  $x \notin T$ . Then,

$$\begin{aligned}
h(\phi_H(T + x)) - h(\phi_H(T)) &= h(\phi_H(T) \cup \phi_H(\{x\})) - h(\phi_H(T)) \\
&\leq h(\phi_H(S) \cup \phi_H(\{x\})) - h(\phi_H(S)) \\
&= h(\phi_H(S + x)) - h(\phi_H(S)),
\end{aligned}$$

where the inequality followed from Inequality (1), and the equalities from the definitions of reachability in a graph. Thus, for any fixed graph  $H$ , the function  $h(\phi_H(T))$  is monotone and submodular in  $T$ . Because the  $\Pr[H]$  are probabilities,  $\mathbb{E}[h(\phi(T))]$  is a non-negative linear combination of monotone submodular functions, and thus also monotone and submodular.  $\blacksquare$

The final piece of the proof of Theorem 3 is the following lemma, showing that monotonicity and submodularity of the Seed Set Model imply diminishing returns for the original model.

**Lemma 16** *Let  $f$  be a non-negative, monotone, submodular function on sets. Consider the function  $g$  defined as follows: Each element  $u$  is included in  $S$  independently with probability  $q_u(\pi_u)$ , where  $q_u$  is an increasing and concave function of  $\pi_u$ . Define  $g(\pi) = \mathbb{E}[f(S)]$ . Then,  $g$  is monotone and satisfies the diminishing returns property as defined in Definition 1.*

**Proof.** First, notice that  $g(\pi) = \sum_{S' \subseteq V} f(S') \prod_{u \in S'} q_u(\pi_u) \prod_{u \notin S'} (1 - q_u(\pi_u))$ . In order to show the diminishing returns property, it is enough to show that  $\frac{\partial g(\pi)}{\partial \pi_i} \geq 0$  and  $\frac{\partial g(\pi)}{\partial \pi_i \partial \pi_j} \leq 0$  for all  $i, j \in V$ . Using the definition of  $g$ , we have:

$$\begin{aligned} \frac{\partial g(\pi)}{\partial \pi_i} &= \sum_{S \subseteq V, i \in S} f(S) \cdot \frac{dq_i(\pi_i)}{d\pi_i} \cdot \prod_{u \in S, u \neq i} q_u(\pi_u) \cdot \prod_{u \notin S} (1 - q_u(\pi_u)) \\ &\quad - \sum_{S \subseteq V, i \notin S} f(S) \frac{dq_i(\pi_i)}{d\pi_i} \cdot \prod_{u \in S} q_u(\pi_u) \cdot \prod_{u \notin S, u \neq i} (1 - q_u(\pi_u)) \\ &= \sum_{S \subseteq V, i \in S} (f(S) - f(S - i)) \cdot \frac{dq_i(\pi_i)}{d\pi_i} \cdot \prod_{u \in S, u \neq i} q_u(\pi_u) \cdot \prod_{u \notin S} (1 - q_u(\pi_u)) \\ &\geq 0. \end{aligned}$$

The last inequality holds because  $\frac{dq_i(\pi_i)}{d\pi_i} \geq 0$  and  $f$  is monotone.

Next we need to show that  $\frac{\partial g(\pi)}{\partial \pi_i \partial \pi_j} \leq 0$  for all  $i, j \in V$ . For  $i = j$ , a calculation similar to the one above shows that

$$\frac{\partial^2 g(\pi)}{\partial \pi_i^2} = \sum_{S \subseteq V, i \in S} (f(S) - f(S - i)) \cdot \frac{d^2 q_i(\pi_i)}{d\pi_i^2} \cdot \prod_{u \in S, u \neq i} q_u(\pi_u) \cdot \prod_{u \notin S} (1 - q_u(\pi_u)),$$

which is non-positive because  $f$  is monotone and  $q_i$  is concave.

Finally, suppose that  $i \neq j$ . Using a calculation similar to the one above, we can rewrite  $\frac{\partial g(\pi)}{\partial \pi_i \partial \pi_j}$  as

$$\sum_{S \subseteq V \setminus \{i, j\}} (f(S + i + j) - f(S + i) - f(S + j) + f(S)) \cdot \frac{dq_i(\pi_i)}{d\pi_i} \cdot \frac{dq_j(\pi_j)}{d\pi_j} \cdot \prod_{u \in S} q_u(\pi_u) \cdot \prod_{u \notin S, u \neq i, j} (1 - q_u(\pi_u)),$$

which is non-positive because  $f$  is submodular and  $q_i, q_j$  are concave.  $\blacksquare$

With Theorem 13 and Lemma 16, we can now complete the proof of Theorem 3.

**Proof of Theorem 3.** Consider one node  $u$ . The probability that it becomes active initially is

$$p_u^0 = \text{Prob}[\lambda_u \pi_u \geq C_u] = 1 - \frac{C_u}{\pi_u}.$$

Recall that  $C_u = c_u \cdot \sum_v d_v$ , and  $\pi_u \geq C_u$  in our model, so this number is always non-negative.

Clearly,  $p_u^0$  is also a monotone increasing function of  $\pi_u$ . To verify concavity, we simply take two derivatives: the second derivative is  $\frac{-2C_u}{(\pi_u)^3}$ , and thus non-positive, so  $p_u^0$  is concave.

Now, consider all the nodes  $u$  which did not initially become active. This is equivalent to saying that  $\lambda_u \pi_u \leq C_u$ . But subject to this bound,  $\lambda_u \pi_u$  is uniformly random, so we are in the situation of having an initially active set  $S$ , and for each remaining node  $u$ , the payment is independently and uniformly random in  $[0, C_u]$ . By Theorems 13 and 15, the expected social welfare  $\overline{W}(S)$  is a monotone and submodular function of the seed set  $S$ , so long as  $W$  is submodular in the set of active nodes. We can therefore apply Lemma 16 to  $E[h(\phi(T))]$ , which implies that  $\overline{W}(\pi_1, \dots, \pi_n)$  has the diminishing returns property. ■

Each of the social welfare functions listed in Section 2 can be shown to be monotone and submodular in the set of active nodes by simple calculations. Thus, for all of these objective functions, the total social welfare is a monotone function of the payments with diminishing returns properties.

## 4 Experimental evaluation

In this section, we summarize our observations based on simulations both on synthetic and real-world P2P networks.

We have developed a simulator for the three models described in Section 2.

### 4.1 Simulation model

Given a payment scheme  $\pi$ , we generate random  $\lambda_u$  and compute the number of active (sharing) nodes. We also compute the value of the social welfare according to the utility functions in Section 2.

In addition, we calculate the total payments, and the average payment per active and per serviced node. These numbers are averaged over 1000 iterations, each with different random  $\lambda$ .

**Network topology.** For our evaluation, we consider different network topologies, including two network topologies derived from real-world data sets [13, 20, 21], and a regular two-dimensional grid topology. The real-world data sets are based on measured end-to-end latencies between pairs of servers deployed in the Internet [13]. The MIT King data set [21] is symmetric and measures RTT between each pair among 1740 servers, while the Harvard King data set [20] provides asymmetric median latencies between each pair among 1895 servers. In addition to networks derived from these two data sets, we also consider a regular two-dimensional grid.

We derive the download percentage matrix  $P$  from the latencies by setting  $p_{v,u} = \max(0, 1 - \frac{\Delta_{u \rightarrow v}}{\Gamma})$ , where  $\Delta_{u \rightarrow v}$  is the latency from  $u$  to  $v$ , and  $\Gamma$  is a hard threshold for tolerable latencies. This models the fact that users prefer to download from peers to which they have fast connections, and have a threshold beyond which latency may not be tolerable any more. By varying  $\Gamma$ , we can obtain denser or sparser download network topologies. We will refer to the networks derived from the MIT King data set as *MIT networks*, and those derived from the Harvard King data set as *Harvard networks*.

In addition to networks derived from these two data sets, we also consider a regular two-dimensional grid. We do not report all results for all topologies here. Unless stated otherwise, our observed trends apply to all of these topologies.

**Payment schemes and non-sharing peers.** In our experiments, we consider different payment schemes  $\pi$ , to study the impact of payments on the propagation of sharing behavior. We parameterize the schemes with two parameters  $\alpha, \beta$ , and set  $\pi_u = \alpha \cdot d_u^\beta$ , where  $d_u$  is the degree of node  $u$  in the network defined by the  $p_{u,v}$  values. Thus, the financial utilities are chosen uniformly at random from the interval  $[0, \alpha \cdot d_u^\beta]$ .

We also consider the impact of peers who cannot (or do not want to) share the file at all, regardless of the payment offered. Such peers may still be interested in downloading the file. Their presence can be expected to decrease the sharing behavior in networks, as they will place load on other peers without contributing. We call such nodes “Empty” nodes, and consider the impact of different percentages of Empty nodes on the overall sharing percentage.

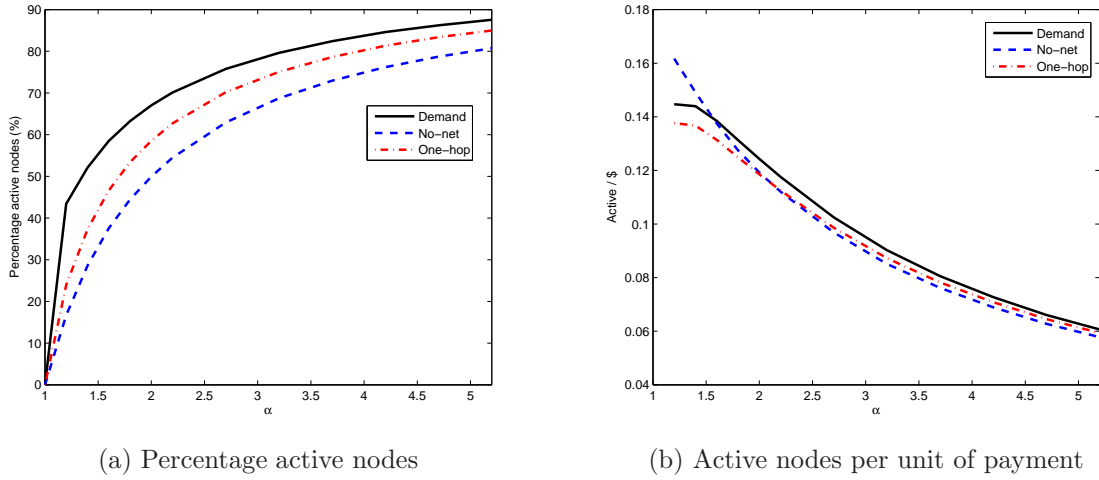


Figure 1: Comparison of different models, using the *Harvard network* with no Empty nodes.

## 4.2 Results

**Comparison of different models.** We begin by estimating the size of network effects, by comparing the *Demand* model with the *No-Network* and *One-Hop* models. Figure 1 (a) compares the participation rates under the three models, with the same payment scheme and same network (*Harvard*). We keep  $\beta = 1$  constant in the payment scheme, and vary  $\alpha$ . Thus, payments are proportional to nodes' degrees. The figure shows that by ignoring network effects, we would underestimate the number of sharing nodes by about 15% on average, and as much as 25% (for  $\alpha = 1.2$ ). The same trends hold for the fraction of *served* nodes (not shown here): the number of *served* nodes is underestimated by about 10% if ignoring network effects.

Figure 1 (b) compares the number of active nodes per unit of payment spent by the network administrator. This is an interesting metric as it captures the tradeoff between participation and payments. Compared to the number of active nodes, the choice of model seems to have remarkably little impact on the estimate of this quantity. For small values of  $\alpha$ , the network effects lead to slightly higher payments per active nodes, as the network effects lead to an activation of more high-degree nodes, which have higher payments. This effect disappears as  $\alpha$  increases, and more nodes are activated in the *No-Network* model as well. The same trends hold for the number of *served* nodes per unit of payment spent (not shown here).

The results reported here stay essentially the same both for the MIT and grid topologies. In particular, the underestimate of the number of active nodes by the *No-Network* model is essentially the same in these topologies. In the grid topology, the *No-Network* model in fact overestimates the cost per active node by about 10%, as the dependence on the degree disappears, and network effects lead to an activation of more nodes with smaller payments.

**Different Social Welfare Functions** We evaluate our theoretical results against the number of *served* nodes and the two social welfare functions sum-welfare and max-welfare, as defined in Section 2. All three are plotted in Figure 2. Although each social welfare function differs from the others in terms of the degree of submodularity (for example, sum-welfare can be shown to be completely modular in the number of active nodes), the curvatures of the plots as a function of payments are more or less the same. Thus, the concavity (diminishing returns) appears to be dominated by the submodularity of the activation process.

**Different payment schemes.** For a network administrator, it is particularly interesting how the choice of payments will affect sharing behavior, and the cost-effectiveness of achieving a certain participation rate. Our next set of experiments therefore shows the percentage of active nodes, and the number of active nodes per unit of payment, when the parameters  $\alpha$  and  $\beta$  in the payments  $\pi_u = \alpha \cdot d_u^\beta$  are varied.

Figure 3 (a) shows the percentage of active nodes in *Harvard network* with 50% Empty nodes, as a



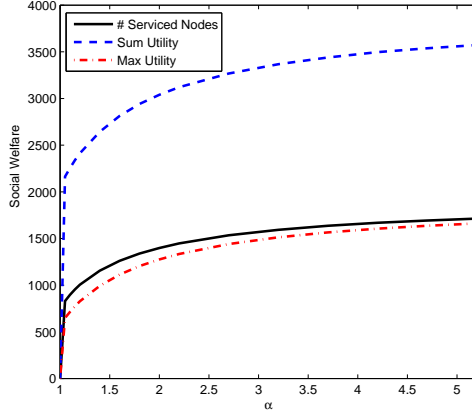


Figure 2: Sum-Welfare, Max-Welfare and the number of serviced nodes, *Harvard network*.

function of  $\alpha$  and  $\beta$ . Figure 3 (b) shows the number of active nodes per unit of payment under the same setting. The cost effectiveness is maximized for very small values of  $\alpha$  and  $\beta$ , specifically  $\beta = 0$  and  $\alpha = 1.2$ . However, this comes at a steep price, in that almost no nodes (only about 4.4% of the network) share in this case.

Clearly, there is no single point at which the network should operate. Rather, a network administrator who wants to achieve a certain participation rate can use these plots find the most cost-effective payment scheme to achieve this rate. For instance, if the goal is to achieve 30% sharing, this can be achieved by setting  $\alpha = 1.6$  and  $\beta = 1.5$ , or  $\alpha = 1.8$  and  $\beta = 1$ . Of these, the first scheme spends about 30 units per active node, while the second scheme spends about 7 units per active node. Thus, a judicious choice of payments can lead to significant savings while ensuring the same level of participation.

In general, the plot suggests that  $\beta \in [0.5, 1]$  tends to lead to good tradeoffs between participation and cost: for smaller values of  $\beta$ , participation tends to be too low, while for higher values, the cost per active node increases significantly.

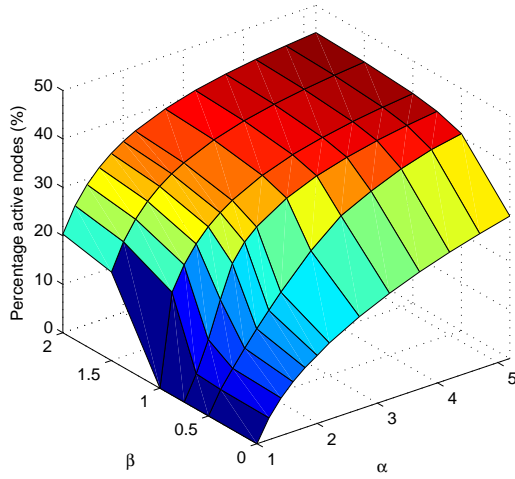
The observed trends are fairly independent of the network topologies. In particular, the plots for both the grid and *MIT network* also suggest that  $\beta \in [0.5, 1]$  gives the best cost efficiency for a given fraction of participating nodes.

**Different thresholds ( $\Gamma$ ).** Finally, we investigate the impact of different latency tolerance thresholds  $\Gamma$  on the activation process. Recall that the larger  $\Gamma$ , the more peers  $u$  may serve  $v$ . For instance, with  $\Gamma = 2\text{ms}$ , the average degree of nodes in the *Harvard network* is 4.58, while with  $\Gamma = 5\text{ms}$ , the average degree increases to 14.93. In the resulting denser graph, we would expect less degree imbalance, and overall higher network effects; however, the payments will need to compensate for more downloads from any individual node.

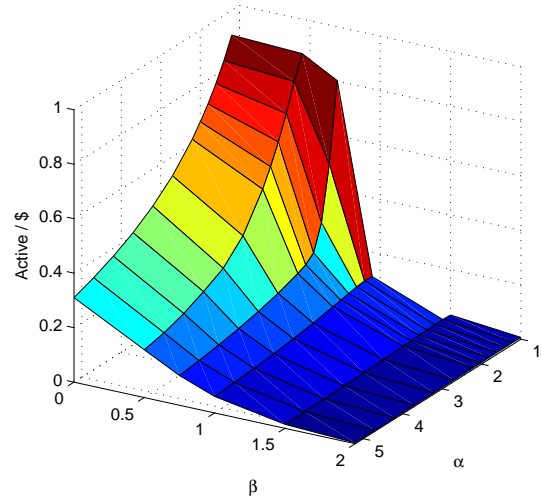
The experiments, conducted on the *Harvard network* with no Empty nodes, confirm this intuition. When  $\beta = 0$ , Figure 4 (a) shows that the number of nodes serviced is smaller in  $\text{Harvard}_{\Gamma=5\text{ms}}$  than in  $\text{Harvard}_{\Gamma=2\text{ms}}$ . The reason is that the payments do not increase with the degree, so it is costlier for nodes in  $\text{Harvard}_{\Gamma=5\text{ms}}$  to become active. As  $\beta$  increases, and high degrees result in higher compensation, more nodes are serviced in  $\text{Harvard}_{\Gamma=5\text{ms}}$ . With  $\beta > 0$ , payments increase in the node degree, and nodes in  $\text{Harvard}_{\Gamma=5\text{ms}}$  receive more payments because of their higher average degree. Thus, more nodes are activated, and as a result, more nodes can be serviced.

The increased activation comes at a price, as seen in Figure 4 (b). The higher average degree in  $\text{Harvard}_{\Gamma=5\text{ms}}$ , combined with the dependence of payments on the degrees, leads to somewhat higher payments per active (or serviced) node. Thus, in the *Demand* model, the increased participation in denser networks is not only a result of network effects, but also of higher payments.

Therefore, in order to investigate the effectiveness of density itself on the participation or service rate, we

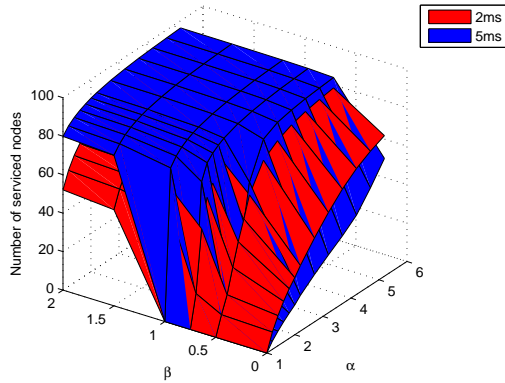


(a) Percentage of nodes active

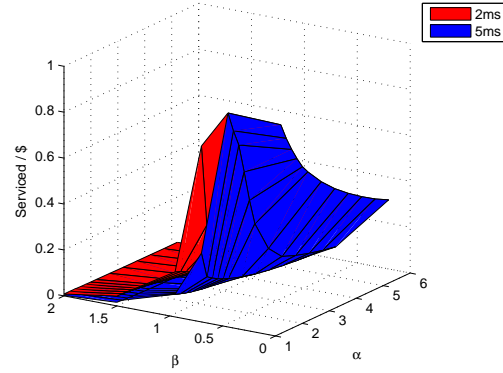


(b) Active nodes per unit of payment

Figure 3: Comparison of different payment schemes, *Harvard network* with 50% Empty nodes. Notice that for readability, the directions of axis labels for  $\alpha, \beta$  is different in the two figures.



(a) Fraction of serviced nodes



(b) Serviced nodes per unit of payment

Figure 4: Comparison of different thresholds, *Harvard network* with no Empty nodes.

make the following comparison. Fix the payment per active node for both  $\text{Harvard}_{\Gamma=5\text{ms}}$  and  $\text{Harvard}_{\Gamma=2\text{ms}}$  to an arbitrary number by choosing the appropriate payment schemes for each graph. For example, in order to get a payment of 23 per active node, a payment scheme for  $\text{Harvard}_{\Gamma=5\text{ms}}$  would be  $\alpha = 2.7$  and  $\beta = 1$  and for  $\text{Harvard}_{\Gamma=2\text{ms}}$  would be  $\alpha = 1$  and  $\beta = 1.5$ . It turns out that the denser network ( $\text{Harvard}_{\Gamma=5\text{ms}}$ ) gives a significantly higher rate for both participation and service. For a payment of 23 units per active node, for instance, the fraction of participating nodes for  $\text{Harvard}_{\Gamma=5\text{ms}}$  is 86% while the same fraction goes down to 39% in  $\text{Harvard}_{\Gamma=2\text{ms}}$ .

Based on the simulations, the following were our main observations:

1. How different are the predictions in sharing behavior between the *Demand*, the *No-Network*, and the

*One-Hop* models? Our results show a significant difference between the models in their prediction of sharing: while the fraction of sharing nodes is qualitatively similar, the predictions ignoring network effects can be off by about 15%–25%. This results in up to 10% depreciation in the number of serviced peers.

2. How does the participation depend on the network topology and density? We observe that the denser the network, the higher the rate of participation, given fixed incentives. This holds across grid and realistic Internet topologies.
3. How does the payment scheme affect the number of sharing and serviced nodes, and the price paid per node? Our experiments suggest that the payments  $\pi_u$  for realistic topologies should be proportional to  $u$ 's degree to give high overall participation at low cost. In other words, given a network topology, there exists a choice of parameters for payments proportional to node degrees that maximizes the overall “bang per buck”. We derive these parameters for each network topology experimentally.

## 5 Conclusions

There are several natural directions for future work. A very interesting question arises when taking payments by “reputation” or download priorities into account. While monetary payments can (in principle) be increased arbitrarily, reputation is inherently constant-sum: if some peers are recognized as outstanding sharers, then others will receive less recognition, and might find the reduced recognition not enough incentive to keep sharing. Similarly, download priorities come at the expense of other peers, and can thus not be arbitrarily increased for all members of the network. As a result, the process of sharing will not necessarily be monotone: peers may choose to stop sharing once too many other peers are active. A first question is then whether stable (equilibrium) states even exist. If so, it would be interesting what fraction of the peers will be sharing, what the social welfare is, and how these quantities will depend on the network structure.

From a more practical viewpoint, it would be desirable to evaluate how accurately our model (or a variation thereof) captures the actual behavior of participants in a P2P system. This would likely be a difficult experiment to perform, as many of the parameters, such as file demands and latency, are inherently transient, and in a realistic system, payments cannot be changed constantly to evaluate the impact of such changes.

In the bigger picture, the network designer also has to be concerned about manipulation by peers. For instance, colluding peers could artificially inflate the perceived “degree” of a peer (by claiming a download preference), and thus the payments to that peer. A more thorough investigation of mechanisms taking these and other concerns into account is an exciting direction for future work.

Finally, our work lies among various applications in economics for which there are positive or negative externalities among agents in a neighborhood. Our results suggest that in order to study different economic metrics such as revenue or social welfare, we should always consider the cascading effect of agents’ strategies over the network.

## References

- [1] Kostas G. Anagnostakis and Michael B. Greenwald. Exchange-based incentive mechanisms for peer-to-peer file sharing. In *Proc. 24th Intl. Conf. on Distributed Computing Systems*, pages 524–533, 2004.
- [2] Christina Aperjis, Michael J. Freedman, and Ramesh Johari. Peer-assisted content distribution with prices. In *Proc. 4th Intl. Conf. on emerging Networking Experiments and Technologies (CONEXT)*, 2008.
- [3] Ravindra B. Bapat. Moore-Penrose inverse of set inclusion matrices. *Linear Algebra and its Applications*, 318(1):35–44, 2000.

- [4] Richard A. Brualdi and Herbert J. Ryser. *Combinatorial Matrix Theory*. Cambridge University Press, 1991.
- [5] Alice Cheng and Eric Friedman. Sybilproof reputation mechanisms. In *Proc. 3rd Workshop on the Economics of Peer-to-Peer Systems (P2PECON)*, 2005.
- [6] Bram Cohen. Incentives build robustness in bittorrent. In *Proc. 1st Workshop on Economics of Peer-to-Peer Systems*, 2003.
- [7] Gérard Cornuéjols, Marshall L. Fisher, and George L. Nemhauser. Location of bank accounts to optimize float. *Management Science*, 23:789–810, 1977.
- [8] Michal Feldman and John Chuang. Overcoming free-riding behavior in peer-to-peer systems. *SIGecom Exchanges*, 5(4):41–50, 2005.
- [9] Michal Feldman, Christos Papadimitriou, John Chuang, and Ion Stoica. Free-riding and whitewashing in peer-to-peer systems. *IEEE Journal on Selected Areas in Communications*, 24(5):1010–1019, 2006.
- [10] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12:211–223, 2001.
- [11] Philippe Golle, Kevin Leyton-Brown, Ilya Mironov, and Mark Lillibridge. Incentives for sharing in peer-to-peer networks. In *Proc. 2nd Intl. Workshop on Electronic Commerce (WELCOM)*, pages 75–87, 2001.
- [12] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83:1420–1443, 1978.
- [13] Krishna P. Gummadi, Stefan Saroiu, and Steven D. Gribble. King: Estimating latency between arbitrary internet end hosts. In *Proc. 2nd Usenix/ACM SIGCOMM Internet Measurement Workshop (IMW)*, 2002.
- [14] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence in a social network. In *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining*, pages 137–146, 2003.
- [15] David Kempe, Jon Kleinberg, and Eva Tardos. Influential nodes in a diffusion model for social networks. In *Proc. 32nd Intl. Colloq. on Automata, Languages and Programming*, pages 1127–1138, 2005.
- [16] Kevin Lai, Michal Feldman, Ion Stoica, and John Chuang. Incentives for cooperation in peer-to-peer networks. In *1st Workshop on Economics of Peer-to-Peer Systems*, 2003.
- [17] Stephen Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.
- [18] Elchanan Mossel and Sebastien Roch. On the submodularity of influence in social networks. In *Proc. 38th ACM Symp. on Theory of Computing*, pages 128–134, 2007.
- [19] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.
- [20] Network Coordinate Research at Harvard.
- [21] Parallel & Distributed Operating Systems Group at MIT.
- [22] Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble. A measurement study of peer-to-peer file sharing systems. In *Proc. SPIE/ACM Conf. on Multimedia Computing and Networking (MMCN)*, 2002.
- [23] Thomas Schelling. *Micromotives and Macrobehavior*. Norton, 1978.

- [24] Jeffrey Shneidman and David C. Parkes. Using redundancy to improve robustness of distributed mechanism implementations. In *Proc. 5th ACM Conf. on Electronic Commerce*, pages 276–277, 2003.
- [25] Vivek Vishnumurthy, Sangeeth Chandrakumar, and Emin Gün Sirer. Karma: A secure economic framework for peer-to-peer resource sharing. In *1st Workshop on Economics of Peer-to-Peer Systems*, 2003.
- [26] Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proc. 39th ACM Symp. on Theory of Computing*, pages 67–74, 2008.
- [27] Ben Q. Zhao, John C. S. Lui, and Dah-Ming Chiu. Analysis of adaptive protocols for p2p networks. In *Proc. 28th IEEE INFOCOM Conference*, pages 325–333, 2009.

## A Hardness of Approximation under the Seed Set Model

Here, we prove that finding a seed set  $S$  to (even approximately) maximize the eventual number of active nodes is hard under the Sharing Process. Let BEST SEED be the optimization problem of finding the seed set  $S$  of at most  $k$  nodes that maximizes the total number of sharing nodes, given  $n$  servers  $u_1, \dots, u_n$  and the corresponding parameters  $c_u, d_u, \lambda_u, \lambda_u \pi_u, p_{v,u}$ . (Notice that when all of the  $\lambda_u$  are given, the process is deterministic.)

**Proposition 17** *It is hard to approximate BEST SEED within  $n^{1-\epsilon}$  for any  $\epsilon > 0$  unless  $P = NP$ .*

**Proof.** We reduce from the VERTEX COVER problem. Recall that the VERTEX COVER problem is formulated as follows: Given a graph  $G = (V, E)$ , a vertex cover is a set  $S \subseteq V$  of nodes such that each edge  $e \in E$  has at least one endpoint in  $S$ . In the VERTEX COVER decision problem, the input is a pair  $(G, k)$ : the question is whether there is a vertex cover of size at most  $k$ . We assume without loss of generality that  $G$  contains no isolated vertices.

Given an arbitrary VERTEX COVER instance with  $N = |V|$  nodes and  $M = |E| \geq N/2$  edges, we construct an instance of BEST SEED as follows: For each node  $u \in V$ , we have a node  $w_u$ . For each edge  $e \in E$ , we create two nodes  $x_e, x'_e$ . Finally, setting  $r = 1/\epsilon$ , we create  $M^r$  “bulk” nodes  $y_1, \dots, y_{M^r}$ . We set  $p_{x'_e, y_i} = 1$  for all  $y_i, x_e$ . For all  $e$ ,  $p_{x'_e, x_e} = 1$ . Finally, whenever  $e$  is incident on  $u$ , we have  $p_{x_e, w_u} = 1$ . All other values of  $p$  are 0.

We visualize the construction above in 4 layers. The “node layer” consists of all nodes  $w_u$  for all  $u$ . The “primary layer” consists of all  $x_e$ . The “secondary layer” consists of all  $x'_e$ . Finally, the “bulk layer” consists of all  $y_i$ . Next, we define payments and demands:

$$\lambda_v \pi_v = \begin{cases} 0 & \text{if } v = w_u \text{ for some } u \in V \text{ (node layer)} \\ 3.5 & \text{if } v = x_e \text{ for some } e \in E \text{ (primary layer)} \\ 0 & \text{if } v = x'_e \text{ for some } e \in E \text{ (secondary layer)} \\ M + 0.5 & \text{otherwise (bulk layer)} \end{cases}$$

$$d_v = \begin{cases} 0 & \text{if } v = w_u \text{ for some } u \in V \text{ (node layer)} \\ 2 & \text{if } v = x_e \text{ for some } e \in E \text{ (primary layer)} \\ 2 & \text{if } v = x'_e \text{ for some } e \in E \text{ (secondary layer)} \\ 0 & \text{otherwise (bulk layer)} \end{cases}$$

First, let  $T$  be a vertex cover of size at most  $k$ . Consider the effect of starting with the nodes  $w_u, u \in T$  as a seed set. Because  $T$  is a vertex cover, each primary node  $x_e$  now has an active node  $w_u$  with  $p_{x_e, w_u} = 1$ , so that its demand of 2 is split between itself and (at least) one node  $w_u$ . Thus, upon activation, it would face at most a demand of 2 from  $x'_e$  and 1 from itself, whereas its payment is 3.5. Hence, each primary node

will become active in the second round. Once the primary node  $x_e$  is active,  $x'_e$  will split its demand evenly between  $x_e$  and all active bulk nodes. Hence, each bulk node  $y_i$  will see demand at most 1 from each  $x'_e$ , for a total of  $M$ . Since its payment offer is larger,  $y_i$  will become active. Hence, all bulk nodes will be active by round 3, and the total number of active nodes is at least  $M^r + M + k$ .

Conversely, suppose that strictly more than  $M + N$  nodes are active. Because none of the secondary nodes ever become active (since they have a payment offer of 0), this means that at least one bulk node must be active. Let  $y_i$  be the first bulk node to become active, breaking ties arbitrarily. Because no other bulk nodes are active at this time,  $y_i$  must see demand at least 1 from each secondary node  $x'_e$ . And because its payment offer is only  $M + 0.5$ , this means that it cannot see demand 2 from any secondary node — otherwise, the total demand would exceed the payment. This means that for each secondary node  $x'_e$ , the corresponding primary node  $x_e$  must already be active. Without loss of generality, the seed set contained no primary nodes — otherwise, the node  $x_e$  could be replaced by  $w_u$  (where  $u$  is an endpoint of  $e$ ), which would next activate  $x_e$ . Thus,  $x_e$  must have become activated at some point of the process, which can only happen when its total demand is smaller than its payment. Since at that point, only  $x_e$  can serve the demand of  $x'_e$ , this in turn means that  $x_e$ 's own demand must be split between itself and one or more active nodes  $w_u$ . Thus, if  $S$  is the set of initially active nodes in the node layer, then the corresponding vertices of  $G$  must form a vertex cover.

In summary, if there is a vertex cover of size at most  $k$ , then there is a seed set of size at most  $k$  activating at least  $M^r + M + k$  nodes, whereas otherwise, no seed set of size at most  $k$  can activate more than  $M + N \leq 3M$  nodes. Thus, no approximation better than  $\Omega(M^{r-1})$  is possible. Since the total number of nodes is  $n = M^r + 2M + N \leq 2M^r$  (for  $r$  large enough), this proves an approximation hardness of  $\Omega(n^{1-1/r}) = \Omega(n^{1-\epsilon})$ , unless  $P=NP$ . ■